II. INTERACTIONS WITH SUMEX-AIM RESOURCE

A. Collaborations and Medical Use of Programs via SUMEX

We have had a modest level of collaboration with a group of students and staff at the Rutgers resource, as well as occasional collaboration with individuals at other ARPA net sites.

B. Sharing and Interactions with Other SUMEX-AIM Projects

As described above, we have had moderate levels of interaction with other members of the SUMEX-AIM community, in the form of writing and reviewing Handbook material. During the development of this material, limited arrangements have been made for sharing the emerging text. As final manuscripts are produced, they will be made available to the SUMEX-AIM community both as on-line files and in the hardcopy, published edition.

C. Critique of Resource Management

Our requests of the SUMEX management and systems staff, requests for additional file space, directories, systems support, or program changes, have been answered promptly, courteously and competently, on every occasion.

III. RESEARCH PLANS (8/80 - 7/83)

A. Long Range Project Goals

The following is the schedule for completion and publication of the AI Handbook:

May, 1981: Publication of Volume 1 by publisher (Wm. Kaufmann Inc., Los Altos, Ca.)

August, 1981: Submission of final copy to publisher for Volume II (publication by end of 1981).

August-September, 1981: Completion of Technical Reports containing chapters of Handbook

October, 1981: Submission of final copy of Volume III to publisher (for publication first quarter 1982)

(note: Volume I has been selected by the Library of Computer Science as their August,1981 book club selection)

B. Justifications and Requirements for Continued SUMEX Use

The AI Handbook Project is a good example of community collaboration using the SUMEX-AIM communication facilities to prepare, review, and disseminate this reference work on AI techniques. The Handbook articles currently exist as computer files at the SUMEX facility. All of our authors and reviewers have access to these files via the network facilities and use the document-editing and formatting programs available at SUMEX. This relatively small investment of resources will result in what we feel will be a seminal publication in the field of AI, of particular value to researchers, like those in the AIM community, who want quick access to AI ideas and techniques for application in other areas.

C. Needs and Plans for Other Computational Resources

We use document preparation programs at SUMEX and the Computer Science Department's SCORE machine. We have used and will continue to use a Computer Science Department phototypesetting machine, the Alphatype, to produce the final copy of the AI Handbook. The phototypesetting software called TEX, developed at Stanford, is the vehicle for this production.

D. Recommendations for Future Community and Resource Development

None.

II.A.1.3 DENDRAL Project

The DENDRAL Project
Resource-Related Research: Computers in Chemistry

Prof. Carl Djerassi
Department of Chemistry
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

The DENDRAL Project is a resource-related research project. The resource to which it is related is SUMEX-AIM, which provides DENDRAL its sole computational resource for program development and dissemination to the biomedical community.

A. Project Rationale

The DENDRAL project is concerned with the application of state-of-the-art computational techniques to several aspects of structural chemistry. The overall goals of our research are to develop and apply computational techniques to the procedures of structural analysis of known and unknown organic compounds based on structural information obtained from physical and chemical methods and to place these techniques in the hands of a wide community of collaborators to help them solve questions of structure of important biomolecules. These techniques are embodied in interactive computer programs which place structural analysis under the complete control of the scientist working on his or her own structural problem. Thus, we stress the word assisted when we characterize our research effort as computer-assisted structure elucidation or analysis.

Our principal objective is to extend our existing techniques for computer assistance in the representation and manipulation of chemical structures along two complementary, interdigitated lines. We are developing a comprehensive, interactive system to assist scientists in all phases of structural analysis (SASES, or Semi-Automated Structure Elucidation System) from data interpretation through structure generation to data prediction. This system will act as a computer-based laboratory in which complex structural questions can be posed and answered quickly, thereby conserving time and sample. In a complementary effort we are extending our techniques from the current emphasis on topological, or constitutional, representations of structure to detailed treatment of conformational and configurational stereochemical aspects of structure.

By meeting our objectives we will fill in the "missing link" in computer assistance in structural analysis. Our capabilities for structural analysis based on the three-dimensional nature of molecules is an absolute necessity for relating structural characteristics of molecules to their observed biological, chemical or spectroscopic behavior. These capabilities will represent a quantum leap beyond our current techniques and open new vistas in applications of our programs, both of which will attract new applications among a broad community of structural chemists and biochemists who will have access to our techniques. This access depends entirely on our access to and the continued availability of SUMEX-AIM. These issues are discussed in detail in the subsequent section, Interactions with the SUMEX-AIM Resource.

The primary rationale for our research effort is that structure determination of unknown structures and the relationship of known structures to observed spectroscopic or biological activity are complex and time-consuming tasks. We know from past experience that computer programs can complement the biochemist's knowledge and reasoning power, thereby acting as valuable assistants in solving important biomedical problems. By meeting our objectives we feel strongly that our programs will become essential tools in the repertoire of techniques available to the structural biochemist.

We are currently beginning the second year of our three year grant. This period represents a transition in the sense that we have pushed our research efforts in techniques for spectral interpretation, structure generation (e.g., CONGEN) and spectral prediction to their limits within the confines of topological representations of molecular structure. At this time, these techniques are perceived to be of significant utility in the scientific community as evidenced by our workshops, the demand for the exportable version of CONGEN and the number of persons requesting collaborative or guest access to our programs at Stanford (see Interactions with the SUMEX-AIM Resource). These existing techniques will, for some years to come, remain as important first steps in solving structural problems. However, in order to anticipate the future needs of the community for programs which are more generally applicable to biological structure problems and more easily accessible we must address squarely the limitations inherent in existing approaches and search for ways to solve them. Our major objectives are based on the following rationale.

None of our techniques (or the techniques of any other investigators) for computer-assisted structure elucidation of unknown molecular structures make full use of stereochemical information. As existing programs were being developed this limitation was less important. The first step in many structure determinations is to establish the constitution of the structure, or the topological structure, and that is what CONGEN, for example, was designed to accomplish. However, most spectroscopic behavior and certainly most biological activities of molecules are due to their three-dimensional nature. For example, some programs for prediction of the number of resonances observed in 13CMR spectra use the topological symmetry group of a molecule for prediction. However, in reality it is the symmetry group of the stereoisomer that must be used. This group reflects the usually lower symmetry of molecules possessing chiral centers and which generally exist in fewer than the total possible number of conformations. This will increase the number of carbon resonances observed over that predicted by the topological symmetry group alone. More generally, few of the techniques in the area of computer-assisted structure elucidation can be used in accurate prediction of structure/property relationships, whether the properties be spectral resonances or biological activities.

A structure is not, in fact, considered to be established until its configuration, at least, has been determined. Its conformational behavior may then be important to determine its spectroscopic or biological behavior. For these reasons we are emphasizing in our current grant period development of stereochemical extensions to CONGEN, our newly-developed structure generator, GENOA (see References 17, 18), and related programs such as the C-13 Nuclear Magnetic Resonance (NMR) programs (see References 15, 16), including machine representations and manipulations of configuration (see References 1, 10) and conformation (see Reference 19) and constrained generators for both aspects of stereochemistry (see References 6, 9, 11, 12).

None of the existing techniques for computer-assisted structure elucidation of unknown molecules, excepting very recent developments in our own laboratory, are capable of structure generation based on inferred partial structures which may overlap to any extent. Such a capability is a critical element in a computer-based system, such as we propose, for automated inference of substructures and subsequent structure generation based on what is frequently highly redundant structural information including many overlapping part structures. Important elements of our research are concerned with further developments of such a capability for structure generation (the GENOA program, (see Reference 17)).

Given the above tools for structure representation and generation, we can consider new interpretive and predictive techniques for relating spectroscopic data (or other properties) to molecular structure (see References 2, 3, 7, 8, 14, 15, 16). The capability for representation of stereochemistry is required for any comprehensive treatment of: 1) interpretation of spectroscopic data (see References 15, 16); 2) prediction of spectroscopic data (see References 15, 16); 3) induction of rules relating known molecular structures to observed chemical or biological properties (see Reference 19). These elements, taken together, will yield a general system for computer-aided structural analysis (the SASES system) with potential for applications far beyond the specific task of structure elucidation.

Parallel to our program development we have embarked on a concerted effort to extend to the scientific community access to our programs, and critical parts of our research effort are devoted to methods for promoting this resource sharing. Our rationale for this effort is that the techniques must be readily accessible in order to be used, and that development of useful programs can only be accomplished by an extended period of testing and refinement based on results obtained in analysis of a variety of structural problems, analyzed by those scientists actively involved in solutions to those problems. Our efforts in this area are summarized in Section II.A, Scientific Collaboration and Program Dissemination).

B. Medical Relevance and Collaboration

The medical relevance of our research lies in the direct relationship between molecular structure and biological activity. The sciences of chemistry and biochemistry rest on a firm foundation of the past history of well-characterized chemical structures. Indeed, structure elucidation of unknown compounds and the detailed investigation of stereochemical configurations and conformations of known compounds are absolutely essential steps in understanding the physiological role played by structures of demonstrated biological activity. Our research is focussed on providing computational assistance in several areas of structural chemistry and biochemistry, with primary attention directed to those aspects of the problem which are most difficult to solve by strictly manual methods. These aspects include exhaustive and irredundant generation of constitutional isomers, and configurational and conformational stereoisomers under chemical, biological and spectroscopic constraints with a guarantee that no plausible stereoisomer has been overlooked.

Although our programs can be applied to a variety of structural problems, in fact most applications by our group and by our collaborators are in the area of natural products, antibiotics, pheremones and other biomolecules which play important biochemical roles. In discussions of collaborative investigations involved with actual applications of our programs we have always stressed the importance of strong links between the structures under investigation and the importance of such structures to health-related research. This emphasis can be seen by examination of the affiliations of current DENDRAL-related investigators and the brief description of current collaborative efforts in Interactions with the SUMEX-AIM Resource.

C. Highlights of Research Progress

In this section we discuss briefly some major highlights of the past year and research currently in progress.

1. Past Year

1.1 Exportable version of the CONGEN program for computer-assisted structure elucidation. CONGEN is an interactive computer program whose task is to provide to the structural biochemist all chemical structures which are possible candidates for the structure of an unknown chemical compound. Based on this information, experiments can be designed to pinpoint the correct structure, thereby facilitating rapid and unambiguous identification of novel, bioactive chemicals. During the past two years we have completed an exportable version of the CONGEN program and have exported it to a variety of structural analysis laboratories in academic, private and industrial research organizations. CONGEN is being utilized at Stanford and at export sites in the hands of investigators who use it as a tool in solving their own structural problems. We have been exporting versions of CONGEN for about 18 months. The program has been used as an aid in the solution of many new structures and recent results have formed the basis for at least eight formal lectures by users of CONGEN at remote sites.

1.2 Version I of the GENOA program for structure generation with overlapping atoms. GENOA (see Reference 17) is an outgrowth of CONGEN whose purpose is to suggest candidate structures for an unknown based on redundant and ambiguous structural inferences. This program, which utilizes CONGEN as an integral part of the computational procedures, is far simpler to use by the practicing biochemist. This results from GENOA's capability to construct structures based on substructural information obtained from a variety of spectroscopic, chemical and biochemical techniques. The program itself considers the structural implications of each new piece of structural data and automatically ensures that all overlaps are considered, thereby freeing the investigator from concerns about the potential for overlapping, or redundant substructural information. In addition, GENOA is the ideal tool for interfacing to automated procedures for spectral interpretation (see References 14, 15), because the necessity for manual intervention in the assignment of substructures is no longer required as it was for CONGEN.

1.3 Programs for Interpretation and Prediction of Spectral Data. We are actively pursuing several novel approaches to the automated interpretation of spectral data, concentrating on carbon-13 magnetic resonance (CMR), proton magnetic resonance (PMR) and mass spectral (MS) data. These approaches utilize large data bases of correlations between substructural features of a molecule and spectral signatures of such features. Our approaches are unique in that: 1) we can incorporate stereochemical features of substructures into the data bases; and 2) we can use the same data bases for both interpretation and prediction of data.

We have recently reported several new developments in the area of analysis of mass spectral data, including methods for mass spectral data interpretation (see Reference 14) and mass spectrum prediction (see References 3, 7, 8).

For either interpretation or prediction of magnetic resonance data, stereochemical substructure descriptors are absolutely essential. Resonance positions are a strong function of the local environment of a resonating atom, including position in space relative to other neighboring atoms. Descriptors which include the three dimensional relationships among atoms in a substructure are required in order to obtain meaningful correlations. We have recently completed the first phases of development of the data base and associated interpretation and prediction programs for C-13 NMR data (see References 15, 16). This approach uses a structure and substructure representation which incorporates configurational stereochemistry (see Reference 16).

Such data bases can be used to interpret spectral data to obtain substructures to be used in CONGEN and GENOA, the structure generating programs (see References 15, 17). Continued automation of this aspect of structure elucidation will significantly ease the burden on the structural biochemist because the computer-based files are much more comprehensive and easier to use than correlation tables or diffuse literature sources. The same data bases can be used to predict spectral signatures in the context of a set of complete molecular structures. Comparison of predicted and observed spectra allows a rank-ordering of candidates and will be very useful in directing the attention of the investigator to the most plausible alternatives (see References 7, 8, 15).

1.4 Constrained generation of configurational stereoisomers. During the previous grant period we solved the problem of computer generation of configurational stereoisomers. These are isomeric chemical structures that differ from one another in the arrangement of atoms in three-dimensional space. We have developed this method further, including now the capability for construction of all possible stereoisomers under stereochemical constraints (see Reference 9). Previously, CONGEN and GENOA were capable only of generation of constitutional isomers which convey no information about the structure in three dimensions. The interaction of biomolecules with biochemical systems is based on their three dimensional nature, not simply their constitution. Therefore, these new developments are crucial to use of computational techniques in structural studies.

Now, for the first time, a computer program can be used to begin with the molecular formula of an unknown compound and using constraints on both molecular connectivity and configuration arrive at a set of structural alternatives which include potential stereochemical variability. This capability allows use of spectral data whose interpretation depends strongly on stereochemical features of molecules. Most importantly, it gives us a structural representation and methods for structure generation and manipulation which represent the foundations for future developments of the one important remaining aspect of structural analysis, treatment of molecular conformations.

2. Research in Progress

The following are some highlights of research in progress. The common theme of these studies is representation of stereochemistry and use of stereochemical information in answering questions concerning the nature of known or unknown molecular structures.

2.1 Development of GENOA and STRCHK. GENOA can now deal with representations of configurational stereochemistry, although it does not make active use of such representations in generating constitutional isomers. The STRCHK (for Structure Checking) program represents the next stage in development of the post-generation analysis programs. STRCHK provides the entry point into the STEREO program for constrained generation of stereoisomers (see Reference 9) from constitutional isomers generated by either CONGEN or GENOA. In the case of GENOA, stereochemical information is passed to STRCHK where it can be used in STEREO. In addition, the mass and C-13 spectrum prediction and ranking programs (see References 7, 8, 14, 15) are available from STRCHK, together with several other utility programs for examining structural candidates. Both programs will be developed further, to the point where export to other computer facilities, as was done with CONGEN, will be possible.

2.2 Development of the C-13 data base and interpretive program. We plan further expansion of the C-13 NMR data base, using data obtained by us from the literature and supplied by others in collaborative efforts. Eventually we would like to pass this work on to an organization better equipped to build and maintain data bases. For the time being, however, our work is sufficiently experimental that we will maintain responsibility for the data base. The C-13 spectrum interpretation program will continue to be developed, as we attempt to make the program more "intelligent" chemically.

2.3 Representation and manipulation of conformational stereochemistry. The next year will see intensive efforts to develop programs for representation of molecular conformations. Preliminary work has led to an algorithm for representation and enumeration of conformations, and to a method for searching for common three-dimensional substructures in a set of structures (see Reference 19). The former study will first be directed to ward a program for representation of substructures with conformation designations. This will lead directly to a method for development of a data base and prediction program for any spectroscopic technique, such as proton NMR, where the spectral signatures are strongly influenced by molecular configurations. Subsequently, a program for generation and, eventually, constrained generation of molecular conformations, will be developed. Parallel to this work, the program for searching for three-dimensional common substructures, a problem important in structure/biological activity correlations, will be developed and tested extensively on previous studies presented in the literature. This work (see Reference 19) is based to an extent on similar work carried out for constitutional representations of structure (see Reference 5).

D. List of Recent Publications

(1) J.G. Nourse, R.E. Carhart, D.H. Smith, and C. Djerassi, "Exhaustive Generation of Stereoisomers for Structure Elucidation," J. Am. Chem. Soc., 101, 1216 (1979).

(2) C. Djerassi, D.H. Smith, and T.H. Varkony, "A Novel Role of Computers in the Natural Products Field," Naturwiss., 66, 9 (1979).

(3) N.A.B. Gray, D.H. Smith, T.H. Varkony, R.E. Carhart and B.G. Buchanan, "Use of a Computer to Identify Unknown Compounds. The Automation of Scientific Inference," Chapter 7 in "Biomedical Applications of Mass Spectrometry, First Supplementary Volume," G.R. Waller and O.C. Dermer, Eds., John Wiley and Sons, Inc., New York, 1980, p. 125.

(4) T.C. Rindfleisch, D.H. Smith, W.J. Yeager, M.W. Achenbach, and A. Wegmann, "Mass Spectrometer Data Acquisition and Processing Systems," in Chapter 3 of "Biomedical Applications of Mass Spectrometry, First Supplementary Volume," G.R. Waller and O.C. Dermer, Eds., John Wiley and Sons, Inc., New York, 1980, p. 55.

(5) T.H. Varkony, Y. Shiloach, and D.H. Smith, "Computer-Assisted Examination of Chemical Compounds for Structural Similarities," J. Chem. Inf. Comp. Sci., 19, 104 (1979).

(6) J.G. Nourse and D.H. Smith, "Nonnumerical Mathematical Methods in the Problem of Stereoisomer Generation," Match, (No. 6), 259 (1979).

(7) N.A.B. Gray, R.E. Carhart, A. Lavanchy, D.H. Smith, T. Varkony, B.G. Buchanan, W.C. White, and L. Creary, "Computerized Mass Spectrum Prediction and Ranking," Anal. Chem., 52 1095 (1980).

(8) A. Lavanchy, T. Varkony, D.H. Smith, N.A.B. Gray, W.C. White, R.E. Carhart, B.G. Buchanan, and C. Djerassi, "Rule-Based Mass Spectrum Prediction and Ranking: Applications to Structure Elucidation of Novel Marine Sterols," Org. Mass Spectrom., 15 355 (1980).

(9) J.G. Nourse, D.H. Smith, and C. Djerassi, "Computer-Assisted Elucidation of Molecular Structure with Stereochemistry," J. Am. Chem. Soc., 102, 6289 (1980).

(10) J.G. Nourse, "Applications of Artificial Intelligence for Chemical Inference. 28. The Configuration Symmetry Group and Its Application to Stereoisomer Generation, Specification, and Enumeration.", J. Amer. Chem. Soc., 101, 1210, (1979).

(11) J.G. Nourse, "Application of the Permutation Group to Stereoisomer Generation for Computer Assisted Structure Elucidation.", in "The Permutation Group in Physics and Chemistry", Lecture Notes in Chemistry, Vol. 12, Springer-Verlag, New York, (1979), p. 19.

(12) J.G. Nourse, "Applications of the Permutation Group in Dynamic Stereochemistry" in "The Permutation Group in Physics and Chemistry", Lecture Notes in Chemistry, Vol. 12, Springer-Verlag, New York, (1979), p. 28.

(13) J.G. Nourse, "Selfinverse and Nonselfinverse Degenerate Isomerizations," J. Am. Chem. Soc., in press (1980).

(14) N.A.B. Gray, A. Buchs, D.H. Smith, and C. Djerassi, "Computer-Assisted Structural Interpretation of Mass Spectral Data," Helv. Chim. Acta, in press (1981).

(15) N.A.B. Gray, C.W. Crandell, J.G. Nourse, D.H. Smith, and C. Djerassi, "Computer-Assisted Interpretation of C-13 Spectral Data," J. Org. Chem., 46 703 (1981).

(16) N.A.B. Gray, J.G. Nourse, C.W. Crandell, D.H. Smith, and C. Djerassi, "Stereochemical Substructure Codes for C-13 Spectral Analysis," Org. Magn. Res., 15, 375 (1981).

(17) R.E. Carhart, D.H. Smith, N.A.B. Gray, J.G. Nourse, and C. Djerassi, "GENOA: A Computer Program for Structure Elucidation Based on Overlapping and Alternative Substructures," j. Org. Chem, 46, 1708 (1981).

(18) D.H. Smith, N.A.B. Gray, J.G. Nourse, and C.W. Crandell, "The DENDRAL PROJECT: Recent Advances in Computer Assisted Structure Elucidation," Anal. Chim. Acta, Computer Techniques and Optimization, in press, (1981).

(19) D.H. Smith, J.G. Nourse, and C.W. Crandell, "Computer Techniques for Representation of Three-Dimensional Substructures and Exploration of Potential Pharmacophores," Proceedings of a Chemical Industries Institute of Technology Symposium on "Structure Activity Correlation as a Predictive Tool in Toxicology," Feb. 10-12, 1981, Raleigh, NC, in press.

E. Funding Support

Title:

RESOURCE RELATED RESEARCH: COMPUTERS IN CHEMISTRY (grant)

Principal Investigator:

Carl Djerassi, Professor of Chemistry, Department of Chemistry, Stanford University

Dennis H. Smith (Associate Investigator), Senior Research Associate, Department of Chemistry, Stanford University

Funding Agency:

Biotechnology Resources Program, Division of Research Resources, National Institutes of Health

Grant Identification Number:

RR-00612-12

Total Award and Period:

Total - 5/1/80 - 4/30/83 --------- $641,419

Current Award and Period:

Current - 5/1/81 - 4/30/82 -------- $237,387

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

In the coming period of our research, our computational approaches to structural biochemistry will become much more general and we plan wide dissemination of the programs resulting from our work. These more general approaches to aids for the structural biochemist will yield computer programs with much wider applicability than, for example, the existing CONGEN, GENOA, STEREO and STRCHK programs. We expect that this will create a significant increase in requests for access to our programs, placing heavy emphasis on our relationship with SUMEX to provide this access (see Justification and Requirements for Continued SUMEX Use for additional details).

For these reasons, in our current grant period the SUMEX-AIM resource is identified as the resource to which our research is related. The SUMEX-AIM resource has provided the computational basis for our past program developments and for initial exposure of the scientific community to these programs. The resource is, however, funded completely separately from our own research; we are only one of a nationwide community of users of the SUMEX-AIM facility. Our relationship to SUMEX is one which goes far beyond mere consumption of cycles on the SUMEX machine. It has been the goal of the SUMEX project to provide a computational resource for research in symbolic computational procedures applied to health-related problems. As such research matures, it produces results, among which are computer programs, of potential utility to a broad community of scientists. A second goal of SUMEX has been to promote dissemination of useful results to that community, in part by providing network access to programs running on the SUMEX-AIM facility during their development phases. SUMEX does not, however, have the capacity to support extensive operational use of such programs. It was expected from the beginning that user projects would develop alternative computing resources as operational demands for their programs grew. Such a state has been reached for the CONGEN, GENOA, STEREO and STRCHK programs and future developments in the DENDRAL Project to yield more generally useful programs will simply magnify the problem.

We will, therefore, under our relationship with SUMEX-AIM, participate as before in the SUMEX-AIM community in sharing methods and results with other groups during development of new programs. In addition, we plan to utilize the small machines requested as part of the SUMEX renewal. Our project will benefit by being able to provide more extensive operational access to our existing and developing programs using these machines, and to provide a test environment for adapting our programs to a more realistic laboratory computing environment than the special-purpose SUMEX resource (see Justification and Requirements for Continued SUMEX Use for additional information). SUMEX will benefit by moving a substantial part of the DENDRAL production load to more cost-effective systems, thereby freeing the SUMEX resource for new program development. Collaborators who wish to use existing programs for specific problems would access SUMEX via the network as before, but now would be routed to new machines. New program developments will be carried out on SUMEX itself, taking advantage of the much more extensive repertoire of peripheral devices, languages, debugging tools and text editors, i.e., precisely the tasks for which that system was designed.

Our proposed relationship to SUMEX-AIM has important implications beyond the practical considerations mentioned above. There is a significant research component to our proposal to make small machines as integral part of the resource sharing aspects of our relationship to SUMEX. The DENDRAL project is one of the first of the SUMEX-AIM projects to have developed sufficient maturity to require additional computer facilities to support production use and to facilitate export of its programs to be applied to real-world, biomedical structural problems. In a sense, then, we will be acting in a pathfinding role for the rest of the SUMEX-AIM community as other projects reach maturity and seek realistic mechanisms for dissemination of their software to meet the computational needs of their collaborators. Cooperating with SUMEX in the use of small machines, implementing new software, regulating access to divert development and applications to the appropriate machine are all experiments which we are willing to undertake together with SUMEX, knowing that we will be providing direction to future efforts along similar lines.

We will also be in a pathfinding role for a large segment of the biochemical community involved in computing, as we explore the utility of machines which will be much more widely available in Department and laboratory environments than DEC-10's and -20's. There are currently very few widely available computing resources which provide access to symbolic, problem solving programs operating in an interactive environment, We would be able to fulfill that need to the extent that applications have direct biomedical relevance, to the limits of our share of the SUMEX-AIM computing resource.

A. Scientific Collaboration and Program Dissemination

Scientific Collaborations:

The following is a brief description of collaborative efforts that have been taking place or will soon commence in the use of DENDRAL programs for various aspects of structural analysis.

1) Drs. Larry Anderson and Elliott Organick, Depts. of Fuels Engineering and Computer Science, University of Utah.

Dr. Anderson's research is in establishing the structure of coal and related polymers via various thermal and chemical degradation schemes. The degradation products are of interest to both energy and environmental studies. Professor Organick is responsible in part for the computer and graphics facility on which CONGEN and related programs can be run. We are exploring with them structure representations based on the Superatom concept in CONGEN as a means of representing families of structures. Access to our programs is primarily via the computer facility at Utah.

2) Dr. Raymond Carhart, Lederle Laboratories.

Dr. Carhart (a former member of our group) is engaged in research concerned with computer applications to structure/activity relationships. Program development is done jointly between Lederle and Stanford with free exchange of software. Lederle applications are carried out on their own computer facility.

3) Dr. Janet Finer-Moore, University of Georgia.

Dr. Finer-Moore is engaged in structure analysis of alkaloids in Dr. Peletier's group at Georgia. This research makes extensive use of 13C NMR. Our collaboration involves the development and application of our 13C interpretive and predictive programs in structure elucidation of new compounds based on an extensive set of 13C data available on closely related compounds. Access is via network to our programs at Stanford. We have just completed the draft of a manuscript as a result of this collaboration. (Dr. Finer-Moore has recently moved to the University of California, San Francisco.)

4) Dr. Brenda Kimble, University of California, Davis.

Dr. Kimble's research is in structural analysis of compounds which are present in trace amounts in environmental milieus and which show mutagenic activity. Many of these compounds are largely aromatic. We are developing the capabilities of our programs to deal efficiently with large, polynuclear aromatic compounds. Access to our programs is via network to Stanford.

5) Dr. Fred McLafferty, Cornell University.

Dr. McLafferty's research is involved with instrumental and analytical aspects of mass spectrometry. We are working with him on the development and application of an interface between his STIRS system and CONGEN/GENOA for structure determination based on mass spectral data. Part of this collaboration is development of IBM versions of some of our programs. Access is in part to Stanford, shifting primarily to Cornell as development proceeds.

6) Dr. David Cowburn, The Rockefeller University

Dr. Cowburn's research is in the area of conformational analysis, primarily of peptides. We are working with him on the development and application of our programs for generation of molecular conformations. Dr. Cowburn's works with large ring peptides which represent a significant challenge for a conformation generator. His participation will help assure an eventual program of practical use rather than just theoretical interest. Collaboration will be via network access to our programs at Stanford.

7) Dr. Gilda Loew, SRI International and The Rockefeller University.

Dr. Loew's research is in the area of quantitative structure/activity relationships, using primarily the methods of quantum mechanical calculations. We are working with her to interface our conformational generator to her coordinate-based calculation methods. Collaboration is carried out via accounts at Stanford with concurrent development of her programs on a VAX facility (NASA Ames Research Center).

8) Dr. D.C. Rohrer, Medical Foundation of Buffalo Research Laboratories, Buffalo, New York.

We have initiated a collaboration with Dr. Rohrer on the problem of finding the common 3-dimensional substructural features of a set of chemical structures. The use of such a program would be to postulate substructural features which are responsible for similar biological or spectral properties. The initial approach is similar to that used successfully to find the greatest common subgraph of a set of constitutional structures. Collaboration will be via network access to Stanford.

9) Dr. J.N. Shoolery, Varian Associates, Palo Alto

We are collaborating with Dr. Shoolery and others at Varian to obtain high quality C-13 spectra of several marine sterols available only in very small quantities. This is being done as part of our ongoing project to develop programs which are capable of spectral interpretation and prediction. The Varian people access our programs directly or via network.

Program Dissemination:

We have provided access to our programs to a community of collaborators via 1) distribution of the CONGEN program to other laboratories, and 2) guest or individual accounts on the SUMEX computer facility here at Stanford. These methods to promote the dissemination and use of our programs are elaborated below, followed by a brief description of some of our collaborations.

a) Program Export

The past two years we have distributed CONGEN to a number of laboratories owning computers on which the exportable version can now execute. These currently include DEC PDP-10 and -20 systems operating under the TENEX, TOPS-10 and TOPS-20 operating systems, and more recently, the beginnings of a version for IBM systems. The following persons are currently running CONGEN on their own laboratory computers:

Dr. Larry Anderson - University of Utah
(work described in section on collaborations)

Dr. Hartmut Braun - Organische-Chemisches Institut der
Universitat Zurich, Switzerland
A former member of Prof. Wipke's group at UC Santa Cruz. He has only recently installed the program at ETH, Zurich.

Dr. Raymond Carhart - Lederle Laboratories
(work described in section on collaborations)

Dr. Roy Carrington - Shell Biosciences Laboratory, England .
Dr. Carrington has used the program both as a guest user and recently in export. He has given presentations on the use of CONGEN and has applied the program to the structure determination of a new acidic amino acid, 2,4-methanoglutamic acid, and other compounds from plant seeds. This work was done in collaboration with Prof. Jon Clardy at Cornell who is also a guest user.

Dr. Robert Carter - University of Lund, Sweden
Dr. Carter obtained a version of the program for use of several groups at Universities in Sweden.

Dr. Daniel Chodosh - Smith, Kline & French Laboratories
He has installed CONGEN and written an extensive users' manual for the use of SKF chemists.

Dr. Henry Dayringer - Monsanto Agricultural Products Co.
He and Dr. Schwenzer (now at Gulf) were responsible for obtaining and installing CONGEN. Primary use is as an aid to structure elucidation of photoproducts and metabolites of agricultural chemicals.

Dr. Douglas Dorman - Lilly Research Labs
Dr. Dorman has been one of our best users. He attended our 1978 workshop and has given several presentations on the use of CONGEN. He has used the program as an aid in solving a number of structures including some beta-lactam antibiotic derivatives.

Dr Philip Ihrig - Amoco Standard Oil (Indiana)

Dr. Martin Huber - Ciba-Geigy, Switzerland
Dr. Huber is a former member of Prof. Wipke's group at UC Santa Cruz. He has recently received the program and is currently working to interest his coworkers at Ciba in computer assisted structure elucidation.

Dr. Carroll Johnson - Oak Ridge National Laboratory
Dr. Johnson is a long time colleague who spent a year at Stanford in 1976. He is involved with the analytical group at Oak Ridge and is using the program as an analytical aid and as a model for programs he is developing.

Dr. G. Jones - ICI Pharmaceuticals, England
He has installed CONGEN and is currently evaluating its utility for use by analytical chemists at ICI.

Dr. Fred W. McLafferty - Cornell University
(work summarized under collaborations)

Dr. Peter W. Milne - CSIRO Division of Computing Research,
Australia
He contacted us through his association with the Heuristic Programming Project at Stanford. He has acted as the Australian contact for distribution of CONGEN in that country.

Dr. James Morrison - Latrobe University, Australia
(see Milne, above)

Dr. David Pensak - E.I. duPont de Nemours and Company
(see EXODENDRAL account DUPONT, and workshop)

Dr. Joseph SanFilippo - Rutgers University
Dr. SanFillippo is using CONGEN in conjunction with his work on superoxide chemistry and in the evaluation of mass spectral data for environmental samples.

Dr. William Sieber - Sandoz, Ltd., Switzerland
He has installed CONGEN for use by structural chemists at Sandoz. Currently they are evaluating its utility.

Dr. M.D. Sutherland - University of Queensland, Australia
(see Milne, above)

Dr. R.O. Watts - Australian National University
(see Milne, above)

b) EXODENDRAL Account

We reserve a special account on SUMEX for persons interested in access to our programs. Initially, this account was used for anyone desiring access, independent of expected level of use or eventual interest. As the SUMEX system became more heavily loaded a mechanism for guest access was provided and at that point we began to differentiate our users by level of interest. For those desiring merely to try programs we provide guest access (see page 119). If there is interest in continuing collaboration, EXODENDRAL status is given, which provides access to more system facilities and good file management capabilities. The persons who have been active under EXODENDRAL status this year are the following (with the account name followed by the contact person and association):

<BRAEKMAN>
Dr. Jean-Claude Braekman - Universite Libre de Bruxelles, Belgium
He is a former post doctoral fellow in our group, and accesses CONGEN from Belgium for natural products structure elucidation.

<BRAUN>
Dr. Hartmut Braun - Organische-Chemisches Institut der Universitat Zurich, Switzerland
(see section on export)

<CARRINGTON>
Dr. Roy Carrington - Shell Biosciences Laboratory, England
(see section on export)

<COWBURN>
Dr. David Cowburn - The Rockefeller University
(see section on collaborations)

<DORMAN>
Dr. Douglas Dorman - Lilly Research Laboratories
(see section on export)

<DREIDING>
Dr. Andre Dreiding - Organische-Chemisches Institut der
Universitat Zurich, Switzerland

He has used CONGEN and STEREO extensively in structural studies. He has also worked closely with Braun (see section on export under Braun).

<DUPONT>
Dr. Earl Abrahamson - E.I. duPont de Nemours and Company

Dr. Abrahamson and 4 colleagues attended our 1980 workshop. They are attempting to integrate our program into their overall computer software system which includes a wide variety of programs for applications to chemical problems.

<FINER-MOORE>
Dr. Janet Finer-Moore - University of Georgia
(see section on collaborations)

<GASH>
Dr. Kenneth Gash - California State College at Dominguez Hills

<HELLER>
Dr. Steven Heller - Environmental Protection Agency

We are continuing our work with the NIH/EPA Chemical Information System, through Heller, to attempt to find mechanisms for making CONGEN accessible through that system.

<HUBER>
Dr. Martin Huber - Ciba-Geigy, Switzerland
(see section on export)

<MILNE>
Dr. Peter W. Milne - CSIRO Division of Computing Research,
Australia
(see section on export)

<MONSANTO>
Dr. Henry Dayringer - Monsanto Company
(see section on export)

<MWOOD>
Dr. Mark Wood - Rutgers University

<RCARHART>
Dr. Raymond Carhart - Lederle Laboratories
(see section on collaborations)

<ROHRER>
Dr. Douglas C. Rohrer - Medical Foundation of Buffalo
(see section on collaborations)

<ROUSSEL>
Dr. Jean Mathieu - Roussel UCLAF
(see section on guest access under Delaroff)

<SIEBER>
Dr. William Sieber - Sandoz Ltd., Switzerland
(see section on export)

<VARIAN>
Dr. James Shoolery - Varian Associates
(see section on collaborations)

c) GUEST Access

We have provided GUEST access to our programs for those persons desiring occasional access to study a structural problem and for those who wish a "hands-on" introduction to the programs. Persons who have received information about this method of access are listed below (and the names of those who have actually logged in as guests are preceded with an asterisk):

*Dr. Robert Adamski - Alcon Labs

*Dr. A. Bothner-by - Carnegie Mellon University
Dr. Bothner-by has requested access to aid others in the Chemistry Department with structure elucidation work.

*Dr. Reimar Bruening - Institut fur Pharmazeutische Arzneimittellehre der Universitat, West Germany
Dr. Bruening has used the program to aid in his solution of the structure of the alkaloid Cassine. He was a participant in our 1978 workshop and has maintained interest since then. He has given at least one presentation in Germany on our programs.

*Dr. William Brugger - International Flavors and Fragrances
Dr. Brugger is interested in eventually obtaining CONGEN for use at IFF in natural products structure elucidation.

*Dr. Robert Carter - University of Lund, Sweden
(see section on export)

*Dr. Francois Choplin - Institut Le Bel, France

*Dr. Jon Clardy - Cornell University
He has used CONGEN on occasion to determine the potential structural variety for an unknown prior to obtaining the X-ray crystal structure.

Dr. Brian Coleman - Koninklijke/Shell-Laboratorium, Holland

*Dr. Mike Crocco - American Hoechst Corp.

*Dr. V. Delaroff - Roussel UCLAF, France
Dr. Delaroff attended our 1980 workshop. He is in charge of a spectroscopic team which checks structures and suggests structures for unknown compounds with important biological activities. They have been using our programs by remote access to aid these investigations.

*Dr. Dan Dolata - University of California at Santa Cruz
He is one of our contacts with Prof. Wipke's group at UC Santa Cruz.

*Dr. Bruno Frei - Laboratorium f. Organische Chemie, Switzerland

*Dr. Y. Gopichand - University of Oklahoma
He has worked with Prof. Schmitz on the solution of several structures of various marine natural products.

*Dr. John Gordon - Kent State University
Dr. Gordon has been using CONGEN while working at Chemical Abstracts in Columbus, Ohio. He has been using CONGEN to investigate general issues of structure representation.

Dr. Peter Gund - Merck, Sharpe and Dolme Research Labs

*Ms. Wendy Harrison - University of Hawaii at Manoa
Ms. Harrison is a student with Prof. Scheuer at Hawaii. She attended our 1978 workshop and has used the programs occasionally as an aid to structure determination in marine chemistry.

Dr. J. Hartenstein - Goedecke Co., Germany

*Dr. Richard Hogue - University of California at Santa Cruz
He is another contact with Prof. Wipke's group.

Dr. H. Honig - Institut fur Organische-Chemie u.
Organisch-Chemische Technologie, Austria

Dr. Kenneth Houk - Louisiana State University
Dr. Houk has recently moved to Pittsburgh where he hopes to develop closer contact with our group.

Dr. H. Kating - Institut fur Pharmazeutische Biologie, Germany

Dr. Brenda Kimble - University of California at Davis
(see section on collaborations)

Dr. Sydell Lewis - University of California at Berkeley

*Dr. David Lynn - University of Virginia
Dr. Lynn attended our 1978 workshop when he was working with Prof. Nakanishi at Columbia.

*Dr. In Ki Mun - Cornell University
(see section on collaborations under McLafferty)

*Dr. Koji Nakanishi - Columbia University
We have worked with him and his students (see Lynn) on structures of several synthetic and natural products.

*Dr. Suba Neir - Washington University, St. Louis
Dr. Neir used the program to aid in determination of the structure of a mutagen.

Dr. A. Neszmelyi - Central Research Institute for Chemistry of the Hungarian Academy of Sciences

Dr. A.C. Oehlschlager - Simon Fraser University, Canada

*Ms. Connie Oshirio - Lawrence Berkeley Labs

Dr. J.R. Jocelyn Pare - The J.R.J. Pare Establishment for Chemistry Ltd., Canada

Dr. James M. Perry - Worcester Polytechnic Institute, Massachusetts

*Dr. Philip Pfeffer - USDA (Philadelphia)

*Dr. Ned Phillips - University of Florida

*Dr. J.D. Roberts - California Institute of Technology

Dr. Robert Santini - Purdue University

Dr. Norm Stemple - Alcon Labs

Dr. Richard Teeter - Chevron Chemical Co.
We have used CONGEN and the mass spectrum analysis programs to verify the structural assignment of an unknown compound.

*Dr. Babu Venkataraghavan - Lederle Laboratories
(see section on collaborations)

Dr. Stephen Wilson - Indiana University

*Dr. W.T. Wipke - University of California at Santa Cruz
We have worked closely with Prof. Wipke's group for several years on problems of structure representation and manipulation in our complementary areas of computer applications in chemistry.

*Dr. Michael Zippel - Institut fur Biochemie Zentrale Arbeitsgruppe Spectroskopie, Germany
Dr. Zippel used CONGEN to investigate the possible connection with their spectral search system.

d) Industrial Affiliates Program

The high level of interest shown by industrial research laboratories in our programs has always presented us with delicate questions about access to SUMEX-AIM. In the past we have granted access for trials of our programs under the conditions that access is necessarily limited and that the recording mechanisms of our programs be used to ensure that all such trial use be in the public domain. As of April, 1980, we began solicitation of interested industrial organizations to participate in a DENDRAL Project Industrial Affiliates Program. As of May 1, 1981, we have six members. We intend to use this program as a means by which we can offer collaborations with our on-going research to industrial organizations separate from SUMEX-AIM. Although EXODENDRAL accounts to such organizations are used to facilitate communication and sharing of new programs and concepts of interest with the community as a whole, all significant and certainly all proprietary use of our programs will be carried out on their own computational facilities.

e) Program License

We are currently exploring the mechanism of program license to commercial firms as a method for dissemination of well-developed programs, for example CONGEN. This mechanism involves a negotiated agreement between a company and Stanford University for rights to access to and dissemination of identified computer programs. Currently, two companies are negotiating with Stanford. We see this mechanism as serving the function of technology transfer in a very realistic way. We do not, as a research project, have the charter or the resources to do what is essentially final engineering of a program and integration of the program into an existing, larger system. Such "value added" effort is crucial to broad acceptance of a computer-based method. In addition, a participating company would take on the burden of maintenance, documentation and training, freeing our personnel to pursue our research objectives and to bring experimental programs to the level of performance where they, too, can be disseminated by licenses.

B. Interactions with Other SUMEX-AIM Projects

We routinely collaborate with other projects on SUMEX most closely related to our own research. In particular, these collaborations have taken place with the CRYSALIS project, MOLGEN, SECS and have begun with Dr. Carroll Johnson at Oak Ridge.

CRYSALIS is concerned with new approaches to the interpretation of X-ray crystallographic data. X-ray crystallography is another approach to molecular structure elucidation. One of our long-term interests is exploring ways in which CONGEN or GENOA generated structures might be used to guide the search of electron density maps. We are also communicating with Prof. Jon Clardy at Cornell on this problem. It is hoped that having narrowed down the structural possibilities for an unknown using physical and chemical data, the few remaining candidates can be used to guide interpretation of such maps.

Most of the structural problems investigated by MOLGEN involve much larger molecules than the size normally investigated in DENDRAL research. Thus, structural representations involving higher levels of abstraction are of utility in MOLGEN, making our structure manipulation tasks quite different. However, many of the ways in which MOLGEN manipulates its structural representations drew on past experience in DENDRAL in developing algorithms to perform these manipulations.

We collaborate frequently with the SECS project in a number of ways. Although our research efforts are in one sense directed toward opposite ends of work on chemical structures, SECS being devoted to synthesis, DENDRAL being devoted to analysis, the underlying problems of structural manipulation share many common aspects. We have exchanged software where possible, particularly in the area of chemical structure display. We have held several discussions in joint group meetings and at several symposia including the AIM Workshops on common problems, including substructure searching, canonical representations and representation and manipulation of stereochemistry. Persons visiting one laboratory often take the opportunity to visit the other. For example, recent visitors to both laboratories have included Prof. Andre Dreiding, Zurich, Dr. Martin Huber, Basel, and Prof. Robert Carter, Lund.

Dr. Carroll Johnson has collaborated on the CRYSALIS project in the past. More recently he has taken an interest in the use of knowledge-based programs for certain problems in spectral data interpretation. For this reason he is exploring the AGE and EMYCIN systems as frameworks for his program structure, and is involved in discussions with DENDRAL to see where common areas of data interpretation can be identified so that he can draw on our experience and programs. This effort is just beginning at this time; we plan to meet early in May at Stanford to continue discussions.

C. Critique of Resource Management

The SUMEX-AIM environment, including hardware, system software and staff, has proven absolutely ideal for the development and dissemination of DENDRAL programs. The virtual memory operating system has greatly facilitated development of large programs. The emphasis on time-sharing and interactive programs has been essential to us in our development of interactive programs. Our experience with other computer facilities has only emphasized the importance of the SUMEX environment for real-world applications of our programs. To run CONGEN, for example, in a batch computing environment would make no sense whatever because the program (and our other, related programs) is successful in large part because an investigator can closely monitor and control the program as it works toward solution. We have no complaints whatsoever about the computing environment.

We do have, however, significant problems with SUMEX-AIM capacity, both in available computer cycles and on-line file storage. In a sense DENDRAL suffers from its success. The rapid progress made during the last grant period and now continuing into the next period has led to development of many new programs as adjuncts to CONGEN and GENOA and at the same time has inspired many persons in the scientific community to request some form of access to our programs. The net result is that it is often very difficult to carry on at the same time development and collaborations involving applications of our programs to structural problems due to high load average on the system.

The current overcrowding we see on SUMEX creates two major problems for us in the conduct of our research. First, it diminishes productivity as many people compete for the resource; the "time-sharing syndrome" leads to idle, wasted time at the terminal waiting for trivial computations to be completed. Second, the slow response time of the system is an aggravation to an outside investigator who is anxiously trying to solve a structural problem. At some point even the most interested persons will give up, log off the computer and resort to manual methods where possible.

We have taken many steps within our project to try to work around heavy use periods on SUMEX. Our group works a staggered schedule, both in terms of the actual hours worked each day and in terms of what days each week are worked. This results in some problems in intra-group communication, but fortunately the message and other communication systems of SUMEX help alleviate that situation. We try to run all demonstrations on the DEC-2020 to help ease the burden on the dual KI-10 system. We encourage our collaborators to avoid prime-time use of the system when possible.

For these reasons, we strongly support the planned augmentation of the SUMEX-AIM hardware. Any part of our computations which can be shifted to another machine will not only facilitate export of our software but will ease the load on the DEC-10s and make it easier to continue our research. Both will serve to make SUMEX more responsive and our productivity higher.

III. RESEARCH PLANS

A. Project Goals and Plans

Current research efforts were described in highlight form in the first section, Summary of Research Program. In this section we discuss in outline form the major goals of our current grant period (5/1/80 - 4/30/83), with an indication of the progress made to date.

Our goals include the following:

1) Develop SASES (Semi-Automated Structure Elucidation System) as a general system for computer aided structural analysis, utilizing stereochemical structural representations as the fundamental structural description. SASES will represent a computer-based "laboratory" for detailed exploration of structural questions on the computer. It will have as key components the following:

A) Capabilities for interpretation of spectral data which, together with inferences from chemical or other data, would be used for determination of (possibly overlapping) substructures. We have made considerable progress in the areas of mass spectrometry (see References 3, 14) and C-13 NMR spectroscopy (see References 15, 18);

B) The GENOA (structure Generation with Overlapping Atoms) program which will have the capability of exhaustive generation of (topological and stereochemical) structural candidates and include as an essential component the existing CONGEN program. We have developed Version I of GENOA for use by our collaborators (see Reference 17);

C) Capabilities for prediction of spectral (and biological) properties to rank-order candidates on the basis of agreement between predicted and observed properties. Again, we have made considerable progress in mass (see References 3, 7, 8) and C-13 NMR (see References 15, 16, 18) spectroscopy;

2) Develop the GENOA program and integrate it with CONGEN. GENOA will represent the heart of SASES for exploration of structures of unknown compounds, or configurations or conformations of known compounds. GENOA will be a completely general method for construction of structural candidates for an unknown based on redundant, overlapping substructural information, and it will include capabilities for generation of topological and stereochemical (see References 1, 6, 9, 10, 11) isomers;

3) Develop automated approaches to both interpretation and prediction of spectroscopic data, including but not limited to the following spectroscopic techniques:

A) carbon-13 magnetic resonance (13CMR) (see References 15, 16, 18);

B) proton magnetic resonance (1HMR);

C) infrared spectroscopy (IR);

D) mass spectrometry (MS) (see References 3, 7, 8);

E) chiroptical methods including circular dichroism (CD), magnetic circular dichroism (MCD).

The interpretive procedures will yield substructural information, including stereochemical features, which can be used to construct structural candidates using GENOA. We have illustrated this method in recent publications (see References 14, 18). The predictive procedures will be designed to provide approximate but rapid predictions of expected spectroscopic behavior of large numbers of structural candidates, including various conformers of particular structures. Such procedures can be used to rank-order candidates and/or conformers. The predictive procedures will also be designed to provide more detailed predictions of structure/property relationships for known or candidate structures in specific biological applications. These procedures have been illustrated in recent publications (see References 3, 7, 8, 15, 18).